Unravelling the Human Interaction with Generative AI-Based Decision Support in Healthcare: Overreliance vs. Overrule of ChatGPT Output in Diagnostic Processes

Abstract
Artificial intelligence (AI) systems to augment the diagnostic process are increasingly implemented in the field of medicine. However, the consequences of incorrectly overruling or overrelying on the output of the AI and, thereby making an incorrect decision, can be crucial for the health outcome of the patient. In this study, we investigate the effect of the interaction with generative AI versus a human expert on the user's ability to lower these two decision errors in medical diagnostics. In addition, we analyze whether the effect differs between participants who have high or low task expertise. In an online experiment, 158 medical students solve two diagnostic tasks with the support of either a human coach or ChatGPT. Preliminary results show that individuals with the support of a human coach informed themselves significantly more about the patient case before interacting with the coach. In addition, individuals with the support of a human coach more frequently asked questions during the interaction to test hypotheses of the right solution than individuals with the support of ChatGPT.

1. Introduction
An increasing number of recommendation systems based on AI are being implemented in various industries. Also in the field of medicine, AI-based systems are increasingly being used to support the human decision-making process. When comparing the performance of humans with the performance of AI-based systems in such situations, the AI system often outperforms the human (Kodagiannis & Lygouras, 2008; Shen et al., 2019). But even if AI systems can achieve better performance than humans, recommendations from an AI are never correct in 100% of cases. It is crucial that the cases in which the AI provides an incorrect recommendation are detected by the user to achieve the key benefit of decision augmentation, meaning that machine and human jointly perform a task, compensating the weakness of each other, i.e. leveraging the strength of each other (Murray et al., 2021). Otherwise, in the field of medicine, the patient's health can be at risk (Hautz et al., 2019).

There are two situations in which an incorrect decision can be made using a supporting AI system. Both situations arise in the case, when human and AI differ in their judgment. On the one hand, the AI provides a correct recommendation, but the individual follows its own incorrect opinion and consequently makes a wrong decision (i.e., overruling the AI). On the other hand, the individual may have a correct assumption, but the AI provides an incorrect recommendation, and the individual makes the wrong decision by following the incorrect recommendation of the AI (i.e., an overreliance on the AI). These decision errors differ depending on the task expertise of physicians (Jussupow et al., 2021). A physician with low task expertise is more likely to overrely on an incorrect output by the AI, and a physician with high task expertise is more likely to overrule a correct output by the AI with its own incorrect opinion (Jussupow et al., 2021). An explanation as to why the overreliance on incorrect statements by an AI decreases with increasing task expertise could be the inheritance of incorrect AI output into one's own opinion (Vicente & Matute, 2023). If one's own correct opinion is overwritten over time with incorrect output from an AI, the initially differing opinion becomes a conforming one, reducing the decision error. Regarding physicians with high task expertise, an explanation for why they more often follow their own incorrect opinions could be the lack of attention to the AI output (Lebovitz et al., 2022). Experts tend not to use AI-generated recommendations (Lebovitz et al., 2022) which could lead them to rely too much on their own opinion.

Such situations, in which correct and incorrect AI output occur, are also expected when using ChatGPT in the diagnostic process, as ChatGPT has an accuracy of 60.3% when making initial differential diagnoses (Rao et al., 2023), but at the same time, ChatGPT shows potential in augmenting diagnostic decisions (Ferdush et al., 2024). Furthermore, a new aspect arises with generative AI tools, such as ChatGPT, compared to traditional AI systems. A traditional AI system generates a recommendation based on input from the patient record, whereas in generative AI tools, the user is actively involved in the process of generating the recommendation. The user does not just evaluate the system's output but can steer the process through the repeated interaction with the generative AI tool and, with its input, influence the quality of the output. Therefore, the question arises: How does the way a user interacts with generative AI influence their ability to lower overreliance and overrule errors in medical diagnostics? Furthermore, how and why do the interactions with generative AI differ from the interactions with a human coach?

To examine this research question, an experiment is conducted with *N* = 158 fourth-year medical students. In the experiment, students are asked to generate differential diagnoses and a final diagnosis for two patient vignettes. During the task, they can use a chat to ask questions as support. One randomly assigned group of participants has a medical intern as their interaction partner replying in real time, while the other group has ChatGPT (version gpt-4-0613) as their chat partner. Throughout the task, all clicks on patient information, chat interactions, and noted differential diagnoses are logged with timestamps. Preliminary results show that individuals with the medical intern as interaction partner acquire more information about the patient case before starting the interaction with their chat partner. In addition, they more frequently ask their chat partner questions to test hypotheses of differential diagnoses during the interaction, than individuals with ChatGPT as interaction partner.

This study contributes to the research on human-AI interaction by quantitatively investigating how people in general and with different task expertise deal with uncertain output and recommendations from generative AI systems and what influence the interaction with the tool has. In addition, we test underlying reasons that could explain why different task expertise results in varying treatment of AI recommendations.

2. Theoretical background and hypothesis generation

When individuals make a decision with the support of an AI system, four scenarios can occur in total regarding the underlying opinions for the decision. To explain the scenarios, a decision between option A and option B is taken as an example, where option A is the objectively better option.

In the first scenario, the individual and the AI system both think option A is the better option, and therefore, the correct decision is being made. In the second scenario, the individual would opt for option B, but the AI system recommends choosing option A. If the individual decides to follow the AI recommendation, he results in the correct decision, if the individual relies on his opinion, he makes the incorrect decision. In the third scenario, the opinions are again differing, but the individual now would prefer option A, whereas the AI system thinks option B is the right choice. The individual can make the right decision by relying on his own opinion, otherwise, by following the recommendation of the AI, the worse option is chosen. In the fourth scenario, both the individual and the AI system would opt for option B, thereby, the incorrect decision is being taken.

The interesting scenarios to investigate are scenarios two and three because the differing opinions can lead to a different decision being taken through the support of an AI system, including a potential decision augmentation. In the medical context, an incorrect decision can have crucial consequences for the patient (Hautz et al., 2019) and the spread of medical diagnostic errors with 5-15% of patients in the healthcare system being affected, illustrates the importance of focusing on situations in scenarios two and three, where diagnostic errors can be reduced or increased (Berner & Graber, 2008; Newman-Toker et al., 2023; Singh et al., 2014). In one situation, decision errors can be increased by an individual who incorrectly refuses the AI recommendation and, therefore, overrules the recommendation of the AI system with its own incorrect opinion. In the other situation, decision errors can increase if an individual incorrectly follows an erroneous AI recommendation, instead of sticking to his own correct opinion. In this situation, the individual shows an overreliance on the recommendation of the AI system.

A study by Jussupow et al. (2021) found that the overreliance and the overrule error differ depending on the task expertise of the individual physicians. Novice physicians who have a low task expertise, more frequently perform the overreliance error, whereas expert physicians with a high task expertise more frequently make the overrule error (Jussupow et al., 2021).

The results raise the question of how these differences in decision error for varying task expertise emerge, which is important to understand in order to reduce the decision errors. There is, to our knowledge, no study that investigates the explanations as to why individuals with low and high task

expertise more frequently overrely and overrule AI recommendation, respectively. A possible explanation for the decreasing overreliance error with increasing task expertise could be the inheritance of incorrect AI output into one's own opinion (Vicente & Matute, 2023). There is an indication that over time, individuals inherit incorrect AI output and, therefore, approximate their opinion to the opinion of the AI (Vicente & Matute, 2023). With this, the first differing opinions become conforming opinions which could result in a decreasing overreliance error.
Regarding the increase of the overrule error with increasing task expertise, previous research indicates that individuals with high task expertise tend to not use the provided AI recommendation (Lebovitz et al., 2022). Consequently, by not using the AI recommendation, in situations with differing opinions, individuals with high task expertise could tend to allocate less attention to the AI output which could lead to increasing overrule errors. In this study, we test these explanations for the influence of task expertise on the overreliance and overrule of ChatGPT output. Decision error is measured with the accuracy of differential and final diagnoses, incorrect output of ChatGPT refers to incorrect AI output, and attention to AI output is measured with the frequency and duration of individuals' interactions with ChatGPT. Therefore, the following hypotheses are derived:

> H1: Individuals with lower task expertise accept incorrect output from ChatGPT more frequently when generating differential diagnoses and final diagnoses than individuals with higher task expertise.
>
> H2: Individuals with higher task expertise interact less frequently as well as for less time with ChatGPT than individuals with lower task expertise.

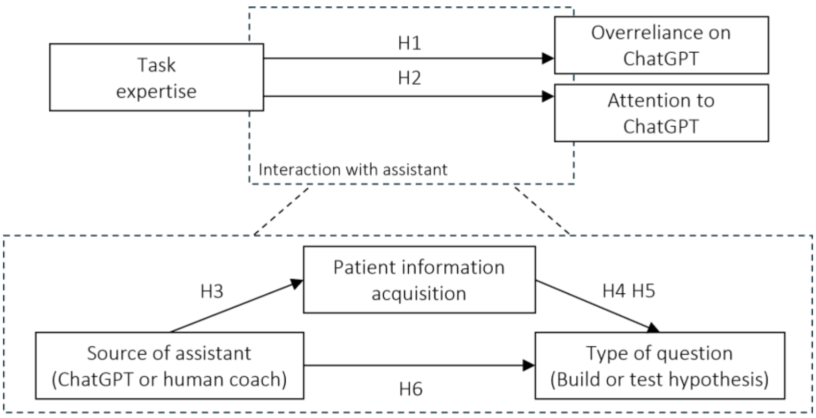

Figure 1 Research model.

The interaction with generative AI systems, such as ChatGPT, has the potential to serve as an alternative to the consultation of advisors or coaches in the diagnostic process to relieve them of their workload which is at its limits due to the skilled labor shortage (Kämmer et al., 2023; Lu et al., 2024; Ten & Durning, 2007). The interaction with ChatGPT resembles a human interaction with generating recommendations through repeated interaction in a dialogue format, however, the interaction itself with the human advisor still differs from the interaction with ChatGPT (Fox & Gambino, 2021). Specifically, the interaction could differ in the following two aspects: the interaction preparation and the question types. Regarding the interaction preparation, since the human aspect is missing in ChatGPT, possible influencing factors, like for example psychological safety (Tynan, 2005), may be less present, and together with the aspect of curiosity towards the AI technology (Huy et al., 2024), leading to more exploratory questions being asked to the ChaGPT assistant in comparison to the human assistant. This exploratory behavior could be observed by how much the individuals inform themselves about a patient case before asking questions in the chat.

H3: Individuals in the ChatGPT condition look at fewer patient information and for less time before they enter the chat than individuals in the condition with a human coach.

Regarding the question types, there are two ways how individuals can collect information through the chat interaction: The assistant can be asked questions to either build hypotheses or to test hypotheses (Schrah et al., 2006). For questions that test hypotheses, an initial assumption about the right solution for the decision is formed and the question to the assistant serves as a support as to whether the assumption should be refused or not (Schrah et al., 2006). When an individual asks a question to build a hypothesis, he does not have an assumption about the right solution and uses the assistant as support to generate assumptions (Schrah et al., 2006). Accordingly, the time and amount of information the individuals look at before asking a question could indicate how the individual will gather information through the interaction in the chat. On the one hand, an individual informing her/himself less about a patient case could lead to asking questions in the chat to build hypotheses. On the other hand, a longer time and a higher amount of information acquisition could lead to asking questions to test hypotheses. Translated to our study, we expect that the amount of patient information acquisition, measured with the duration and the amount of patient information individuals look at, will lead to different question types. Individuals with higher (vs. lower) initial patient information acquisition could more frequently ask questions to test (vs. build) a hypothesis as a first question, as well as throughout the chat interaction. This leads to the following hypotheses:

H4: Individuals who look at fewer patient information and for less time before they enter the chat more frequently ask questions in the chat to build hypotheses.

H5: Individuals who look at a more patient information and for a longer time before they enter the chat more frequently ask questions in the chat to test hypotheses.

H6: During the whole chat interaction, individuals in the ChatGPT condition more frequently ask questions to build hypotheses than individuals in the human coach condition.

An overview of the hypotheses is presented in the research model in Figure 1.

3. Method

To investigate the hypotheses, an online experiment is conducted with medical students from the Charité Medical School in Berlin. Each student receives a financial reimbursement of 35 Euros for his participation in the study. The data collection started on 22 April, 2024 and until 1 August, 2024 the answers of two thirds of the sample have been collected. Due to an insufficient number of students registering to take part in the study, the data collection has not yet been completed, the data collection is expected to finish by the end of August.

3.1 Design

A 2x2 between-subjects factorial design is used with source of assistance and training for the two factors. The factor source of assistance consists of two levels (human coach and ChatGPT), as well as the factor training (training and no training). In the training condition, participants received on the one hand information on diagnostic errors and three important underlying and contributing factors limited knowledge, premature closure and overconfidence (Berner & Graber, 2008; Gäbler, 2017). On the other hand, examples for questions to ask their chat assistant to circumvent the three factors are presented, before the participants started the diagnostic tasks.

The conditions of the two factors are randomly assigned and the participants are blinded for the factor training. However, the students are informed about the source of assistance they receive in the beginning of the experiment.

3.2 Materials and Procedure

3.2.1 First session

The data collection is split up into three sessions (see Figure 2). The medical students who are invited to participate in the study via mailing lists, posters, and online platforms of the Charité Skills Lab, watch a general introduction video of large language models (LLMs) in the first session. A short survey follows the video in which participants sign consent on the participation and answer baseline questions regarding their attitude and experience with AI tools, as well as demographic questions (all original items of the surveys in the three sessions and their English translation can be accessed at https://osf.io/cbpr3/?view_only=e5e94231ddd546b491c2e07f43f02c88).

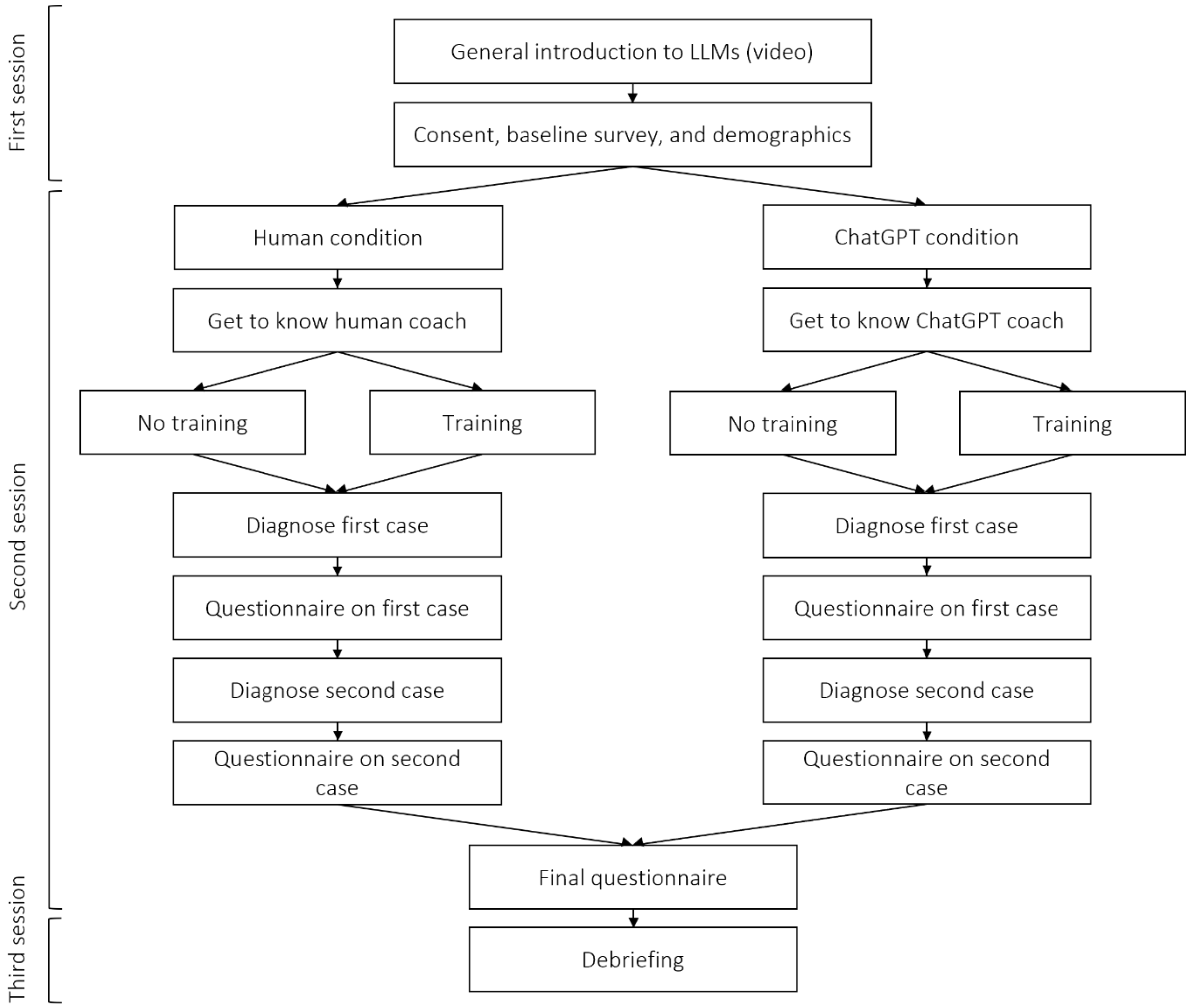

Figure 2 Procedure of the experiment.

3.2.2 Second session

In the second session, up to six participants are invited to an MS Teams meeting. In the beginning of the second session, the participants are introduced to the study, then they are randomly assigned to one of each condition (i.e., either the human coach or ChatGPT condition and either the training or no training condition) and sent to an individual breakout room to take part in the diagnostic tasks.

To start the diagnostic tasks, participants pass a get-to-know phase in which they get to know their respective chat assistant, as well as his strengths and weaknesses. In the next step, participants assigned to the training condition first pass the training, whereas the participants from the no training condition are directly forwarded to the diagnostic tasks.

The diagnostic tasks consist of diagnosing two patient cases that are presented in random order. The patient cases in each of the two diagnostic tasks are based on real emergency cases (Kumar et al., 2011;

Kunina-Habenicht et al., 2015). The equivocal cases have a correct diagnosis as the solution of the task, but also a main incorrect competing diagnosis. In one case, the correct diagnosis is pulmonary embolism with myocardial infarction as the incorrect competing diagnosis, and in the other case, the correct diagnosis is pulmonary embolism with stroke as the incorrect competing diagnosis, respectively.
Participants see information about the patient case (i.e., patient history, blood samples, laboratory results, medical imaging, and ECGs) by clicking on the equivalent tab in the interface (see Figure 3). Below the patient information, participants enter all differential diagnoses they consider during the diagnostic task. By clicking on arrow buttons, they can sort the differential diagnosis and they can also delete them from the list by clicking on the garbage bin button. On the right side of the interface, participants have the possibility to chat with their assigned assistant in real time (i.e., a human coach or ChatGPT) as a support to solve the diagnostic task. Red boxes in the chat window represent text messages from the assistant and blue boxes represent messages from the user. During the diagnostic tasks, all clicks, noted differential diagnoses, and chat interactions are logged with timestamps.
After each diagnostic task, participants rate the likelihood of each of their noted differential diagnoses on a scale from 0 to 100, write a reason for their final diagnosis, and state what next steps they would undertake with the patient in the case. Following the information on the differential and final diagnoses, participants fill in a survey regarding the perception of the difficulty of the case, the familiarity of the case, the competence of the assistant, and the support of the assistant.
In the next step, participants complete a final survey that collects information regarding the perception of the usefulness, satisfaction, and credibility of the assistant. Finally, participants return from the breakout session to the main meeting room and are informed about the debriefing in the third session.

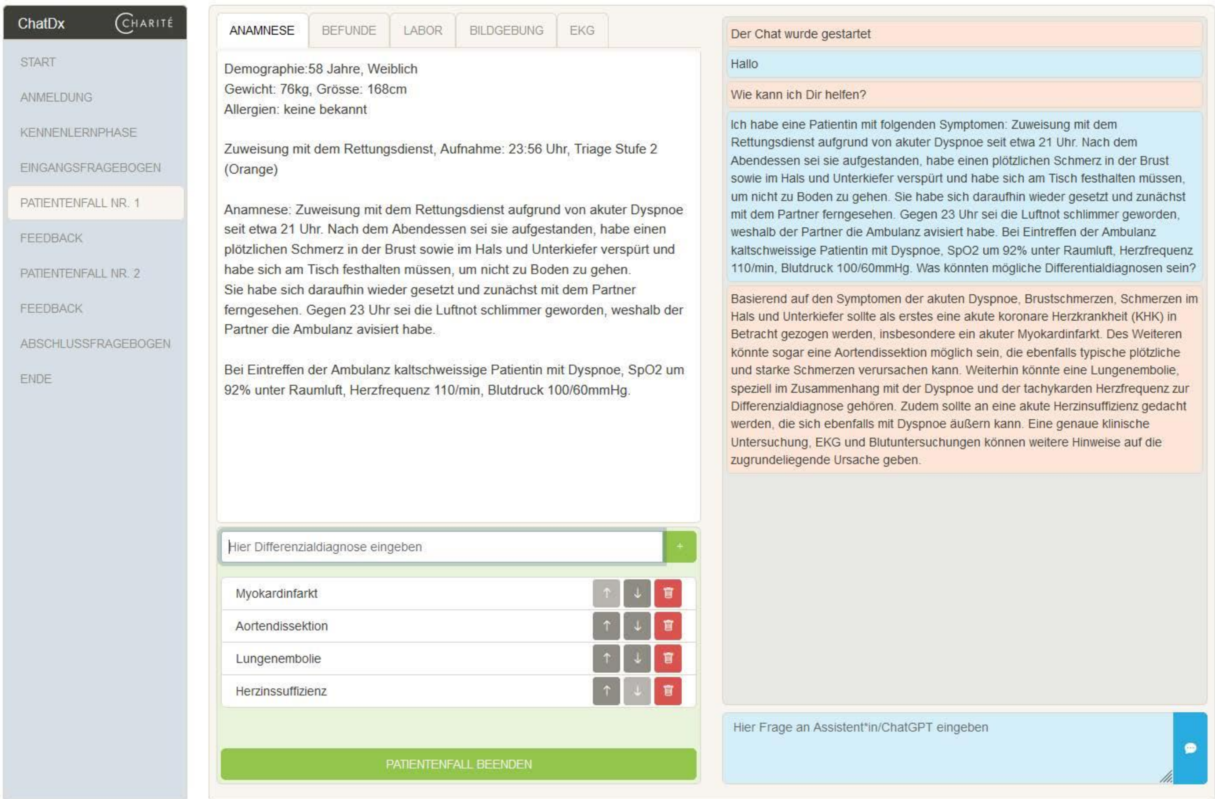

Figure 3 A screenshot of the patient case interface in the diagnostic task.

3.2.3 Third session
After all participants completed the second session, a debriefing is held for all participants in the third session. Participants will receive information regarding the solutions to the diagnostic tasks, the training instructions, and additional resources on clinical decision-making and LLMs.

3.3 Assistant

3.3.1 Human coach

In the human coach condition, the assistant in the chat is a medical intern who works in the hospital. In total, two medical interns are instructed as human coaches. They passed a training of five hours on the philosophy of peer teaching (Ten & Durning, 2007) and deliberate reflection (Mamede et al., 2008). The training also included information on the study purpose and the chat system. The two medical interns are randomly assigned to participants of the human coach condition and to answer frequent questions in a quick and standardized way, they receive scripts with standard answers. To eliminate gender bias and to keep the names of the medical interns confidential, they are called 'Toni', an unisex name, during the chat interaction. The medical interns are reimbursed with 20 Euros per hour for their participation in the study.

3.3.2 ChatGPT

In the ChatGPT condition, participants can chat with ChatGPT (version gpt-4-0613, DeploymentName = 'GPT-4', MaxTokens = 1000, Temperature = 1.0f) as a support during the diagnostic task. ChatGPT is integrated in the chat window of the interface of the diagnostic task with an application programming interface of the Microsoft Azure cloud plaform which is hosted in the data center 'Switzerland North'.

3.4 Sample

The target sample size was calculated to be N = 158 using G*Power 3.1.9.7 (Faul et al., 2007) for a 2x2 analysis of variance to detect a medium effect size with $\alpha = .05$, and $\beta = .80$.
The participation in the study is possible for all medical students at the Charité Medical School in Berlin that are in the fourth year of their studies (N = 640), are at least 18 years old, and have given their written consent for participation.

4. Results

4.1 Pilot study

To test the case materials, we conducted a pilot study with N = 11 participants (Age: M = 26 years, SD = 4.9; Gender: 55% female). In the pilot study, the participants did not have access to an assistant. The results of the pilot study showed that 27% of participants correctly solved the first diagnostic case by stating pulmonary embolism as their final diagnosis. The correct solution to the second diagnostic case (i.e., aortic dissection) was stated by 0% of participants as their final diagnosis. This confirmed the appropriate difficulty of the cases in order to prevent ceiling effects.

4.2. Main study

To analyze the hypotheses, the question types of the users in the chat interactions and the differential diagnoses have to be coded. The accuracy of differential diagnoses is calculated at the end of the data collection through the number of required steps between the International Classification of Diseases (ICD; 10th revision; ICD-10) codes of the correct diagnosis and the ICD codes of the generated diagnoses. Until 1 August 2024, over 130 unique differential diagnoses have been created so far by all participants.
The coding scheme for the question types of the users was developed deductively and adapted inductively with the data of the first ten chat interactions. Afterward, the chat interactions of 20% of the participants are coded by two medical master students using MAXQDA, with multiple meetings organized throughout the coding process to answer questions and complete the coding scheme with examples to more clearly distinguish the different codes of the coding scheme. Based on the coded chat interaction, the intercoder agreement was assessed (see Table 1). Most codes were generated for the question type category 'Verification: Support for checking and excluding differential diagnoses' (80 codes, 33.76% of all codes generated), followed by the category 'Request: Support for generating a differential diagnosis' (70 codes, 29.54% of all codes generated) and 'Statement: Expression of a suspected differential diagnosis' (41 codes, 17.3% of all codes generated). Across all coding categories, coder A and coder B showed an

average intercoder agreement of 74.45%. To ensure intercoder reliability, disagreeing coding segments will be discussed with the coders and afterward, the rest of the chat interactions are randomly divided between the two coders.

Table 1 Intercoder Agreement.

Question types of users	Coder A	Coder B	Agreement[†]	Disagreement	Agreement percentage
Technical questions: Close a knowledge gap	17	17	10	14	41.67%
Request: Support for generating a differential diagnosis	35	35	34	4	89.47%
Statement: Expression of a suspected differential diagnosis	20	21	16	9	64%
Differentiation: Support for differentiating between differential diagnosis	6	6	6	0	100%
Verification: Support for checking and excluding differential diagnoses	41	39	36	8	81.82%
Diagnostics: Support for selecting diagnostics	0	0	0	0	-
Management: Support for the management and the next steps being taken	0	0	0	0	-
Total	119	118	102	35	74.45%

[†] Segments with a minimum code overlapping rate of 90% are counted as an agreement.

Since the data collection is still in progress, the hypotheses cannot be conclusively evaluated. Some variables needed for the analysis of the hypotheses can only be generated after the data collection has been completed. However, the variables used to analyze H3, H4, H5, and H6 can already be generated, and preliminary results can be provided by analyzing the cleaned data of the first 20% of participants. To preliminarily test H3, two linear mixed effects models with random intercepts for participants and tasks were conducted. The dependent variables included duration (measured in seconds) and amount of initial patient information acquisition, the independent variable consisted of the variable source of assistance (dummy-coded as 1 = ChatGPT, and 0 = human coach, respectively). Results show that individuals in the human coach condition looked significantly longer (b = 221.11, p = .00) and at more (b = 0.53, p = .03) patient information before they entered the chat.

For the evaluation of H4 and H5, two logistic mixed effects models with random intercepts for participants and tasks were used with type of question (dummy-coded for H4 as 1 = build hypothesis, 0 = test hypothesis, and for H5, 1 = test hypothesis, 0 = build hypothesis, respectively) as dependent variable and duration as well as amount of initial patient information acquisition as independent variables. Individuals who looked at fewer (vs. more) patient information (b = -0.27, p = .16) and for less (vs. longer) time (b = -9.295e-05, p = .87) did not ask more frequently a question to build (vs. test) a hypothesis in the initial interaction, the null hypothesis of H4 and H5 cannot be rejected.

To analyze H6, the influence of the source of assistance on the types of questions being asked during the whole chat interaction was examined with two linear mixed effects models. The absolute amount of questions being asked to build or to test hypotheses formed the dependent variables, the independent variable included the variable source of assistance (dummy-coded as 1 = ChatGPT, and 0 = human coach, respectively). Results show that individuals in the ChatGPT condition asked less frequently questions to test hypotheses (b = -4.01, p = .01) than individuals in the human coach condition (b = 10.23, p = .01).

To investigate H1 and H2, the final data is planned to be analyzed with mixed effects models, including random effects for tasks and participants. A detailed description of the planned analysis of all hypotheses can be accessed at https://osf.io/g6ncz.

5. Discussion

This study contributes to the research on human-AI interaction by quantitatively investigating how people in general and with different task expertise deal with uncertain output and recommendations from generative AI systems and what influence the interaction with the tool has. In addition, we test underlying reasons that could explain why different task expertise results in varying treatment of AI recommendations. Furthermore, the comprehensive analysis of the chat interactions allows a deeper understanding of the usage and influence of ChatGPT support on the diagnostic process.

Hospitals or professionals in medicine could use the findings of this study for the design of LLM interfaces for physicians and for the development of measurements to reduce diagnostic errors.

Finally, two methodological aspects may be considered in future research. First, as task expertise is a variable that evolves over time, a longitudinal study could investigate the effect of AI support in decision-making or problem-solving tasks over a longer period of time. With this, the dynamics of the usage of the AI support and the reliance on the AI output could be analyzed longitudinally in addition to this cross-sectional study. Second, the study could be conducted using a task whose consequences pose less risk. It can be assumed that when human life is at stake, participants deal differently with output from ChatGPT than for work tasks that involve less risk (e.g., tasks in software development, administration, or education).

References

Berner, E. S., & Graber, M. L. (2008). Overconfidence as a cause of diagnostic error in medicine. *The American Journal of Medicine*, *121*(5), S2-S23.

Faul, F., Erdfelder, E., Lang, A. G., & Buchner, A. (2007). G* Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods, 39*(2), 175-191.

Ferdush, J., Begum, M., & Hossain, S. T. (2024). ChatGPT and clinical decision support: scope, application, and limitations. *Annals of Biomedical Engineering*, *52*(5), 1119-1124.

Fox, J., & Gambino, A. (2021). Relationship development with humanoid social robots: Applying interpersonal theories to human–robot interaction. *Cyberpsychology, Behavior, and Social Networking*, *24*(5), 294-299.

Gäbler, M. (2017). Denkfehler bei diagnostischen Entscheidungen. *Wiener Medizinische Wochenschrift*, *167*(13), 333-342.

Hautz, W. E., Kämmer, J. E., Hautz, S. C., Sauter, T. C., Zwaan, L., Exadaktylos, A. K., ... & Schauber, S. K. (2019). Diagnostic error increases mortality and length of hospital stay in patients presenting through the emergency room. *Scandinavian journal of trauma, resuscitation and emergency medicine*, *27*, 1-12.

Huy, L., Nguyen, H., Vo-Thanh, T., Thinh, N., & Thi Thu Dung, T. (2024). Generative AI, Why, How, and Outcomes: A User Adoption Study. *AIS Transactions on Human-Computer Interaction, 16*(1), 1-27. https://doi.org/10.17705/1thci.00198

Jussupow, E., Spohrer, K., Heinzl, A., & Gawlitza, J. (2021). Augmenting medical diagnosis decisions? An investigation into physicians' decision-making process with artificial intelligence. *Information Systems Research*, *32*(3), 713-735.

Kämmer, J. E., Choshen-Hillel, S., Müller-Trede, J., Black, S. L., & Weibler, J. (2023). A systematic review of empirical studies on advice-based decisions in behavioral and organizational research. *Decision*, *10*(2), 107.

Kodogiannis, V., & Lygouras, J. N. (2008). Neuro-fuzzy classification system for wireless-capsule endoscopic images. *International Journal of Electrical, Computer, and Systems Engineering*, *2*(1), 55-63.

Kumar, B., Kanna, B., & Kumar, S. (2011). The pitfalls of premature closure: clinical decision-making in a case of aortic dissection. *Case Reports*, *2011*, bcr0820114594.

Kunina-Habenicht, O., Hautz, W. E., Knigge, M., Spies, C., & Ahlers, O. (2015). Assessing clinical reasoning (ASCLIRE): Instrument development and validation. *Advances in Health Sciences Education*, *20*, 1205-1224.

Lebovitz, S., Lifshitz-Assaf, H., & Levina, N. (2022). To engage or not to engage with AI for critical judgments: How professionals deal with opacity when using AI for medical diagnosis. *Organization Science*, *33*(1), 126-148.

Lu, Z., Wang, D., & Yin, M. (2024). Does more advice help? the effects of second opinions in AI-assisted decision making. *Proceedings of the ACM on Human-Computer Interaction*, *8*(CSCW1), 1-31.

Mamede, S., Schmidt, H. G., & Penaforte, J. C. (2008). Effects of reflective practice on the accuracy of medical diagnoses. *Medical Education*, *42*(5), 468-475.

Murray, A., Rhymer, J. E. N., & Sirmon, D. G. (2021). Humans and technology: Forms of conjoined agency in organizations. *Academy of Management Review*, *46*(3), 552-571.

Newman-Toker, D. E., Peterson, S. M., Badihian, S., Hassoon, A., Nassery, N., Parizadeh, D., ... & Robinson, K. A. (2023). Diagnostic errors in the emergency department: a systematic review.

Rao, A., Pang, M., Kim, J., Kamineni, M., Lie, W., Prasad, A. K., Landman, A., Dreyer, K. & Succi, M. D. (2023). Assessing the utility of ChatGPT throughout the entire clinical workflow: development and usability study. *Journal of Medical Internet Research*, *25*, e48659.

Schrah, G. E., Dalal, R. S., & Sniezek, J. A. (2006). No decision-maker is an island: Integrating expert advice with information acquisition. *Journal of Behavioral Decision Making*, *19*(1), 43-60.

Shen, J., Zhang, C. J., Jiang, B., Chen, J., Song, J., Liu, Z., He, Z., Wong, S.Y., Fang, P., & Ming, W. K. (2019). Artificial intelligence versus clinicians in disease diagnosis: systematic review. *JMIR Medical Informatics*, *7*(3), e10010.

Singh, H., Meyer, A. N., & Thomas, E. J. (2014). The frequency of diagnostic errors in outpatient care: estimations from three large observational studies involving US adult populations. *BMJ quality & safety*, *23*(9), 727-731.

Ten Cate, O., & Durning, S. (2007). Dimensions and psychology of peer teaching in medical education. *Medical teacher*, *29*(6), 546-552.

Tynan, R. (2005). The effects of threat sensitivity and face giving on dyadic psychological safety and upward communication 1. *Journal of Applied Social Psychology*, *35*(2), 223-247.

Vicente, L., & Matute, H. (2023). Humans inherit artificial intelligence biases. *Scientific Reports*, *13*(1), 15737.